

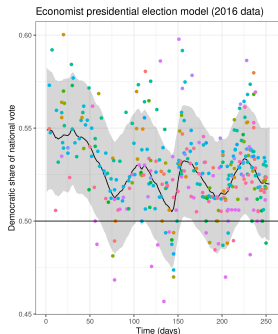
# Approximate data deletion and replication with the Bayesian influence function

---

Ryan Giordano (rgiordano@berkeley.edu, UC Berkeley), Tamara Broderick (MIT)

**MIT Robustness and Influence Functions Workshop**

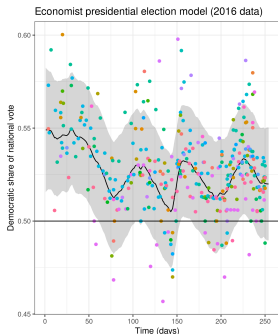
# Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \dots, x_N =$  polling data.
- $\theta =$  parameters for everything we don't know:
  - Daily information shocks (time series residuals)
  - Ideosyncracies of particular polling agencies
  - Biases in polling methods
- Model polling randomness with  $p(X|\theta)$
- $f(\theta) =$  Democratic % of vote on election day

# Economist 2016 Election Model [Gelman and Heidemanns, 2020]



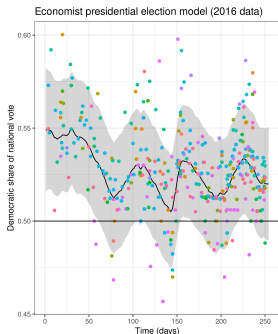
A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \dots, x_N =$  polling data.
- $\theta =$  parameters for everything we don't know:
  - Daily information shocks (time series residuals)
  - Ideosyncracies of particular polling agencies
  - Biases in polling methods
- Model polling randomness with  $p(X|\theta)$
- $f(\theta) =$  Democratic % of vote on election day

Many values of  $\theta$  are consistent with the data.

How can we account for our uncertainty when making election predictions?

# Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

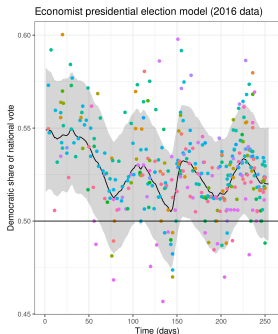
- $X = x_1, \dots, x_N =$  polling data.
- $\theta =$  parameters for everything we don't know:
  - Daily information shocks (time series residuals)
  - Ideosyncracies of particular polling agencies
  - Biases in polling methods
- Model polling randomness with  $p(X|\theta)$
- $f(\theta) =$  Democratic % of vote on election day

Many values of  $\theta$  are consistent with the data.

How can we account for our uncertainty when making election predictions?

**Bayesian answer:** Average over values that are plausibly consistent with the data.

# Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \dots, x_N =$  polling data.
- $\theta =$  parameters for everything we don't know:
  - Daily information shocks (time series residuals)
  - Ideosyncracies of particular polling agencies
  - Biases in polling methods
- Model polling randomness with  $p(X|\theta)$
- $f(\theta) =$  Democratic % of vote on election day

Many values of  $\theta$  are consistent with the data.

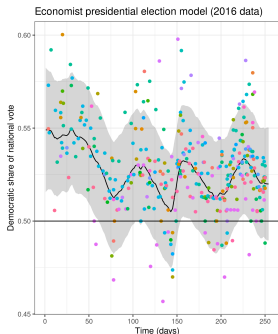
How can we account for our uncertainty when making election predictions?

**Bayesian answer:** Average over values that are plausibly consistent with the data.

**Formally:** Define a prior  $p(\theta)$ , compute the posterior  $p(\theta|X)$ , and estimate

$$f(\theta) \approx \mathbb{E}_{p(\theta|X)} [f(\theta)], \text{ with uncertainty proportional to } \sqrt{\text{Var}_{p(\theta|X)} (f(\theta))}.$$

# Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \dots, x_N =$  polling data.
- $\theta =$  parameters for everything we don't know:
  - Daily information shocks (time series residuals)
  - Ideosyncracies of particular polling agencies
  - Biases in polling methods
- Model polling randomness with  $p(X|\theta)$
- $f(\theta) =$  Democratic % of vote on election day

Many values of  $\theta$  are consistent with the data.

How can we account for our uncertainty when making election predictions?

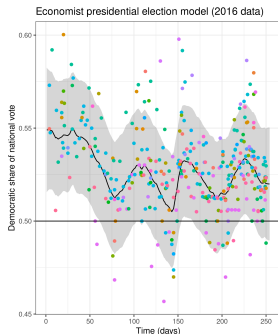
**Bayesian answer:** Average over values that are plausibly consistent with the data.

**Formally:** Define a prior  $p(\theta)$ , compute the posterior  $p(\theta|X)$ , and estimate

$$f(\theta) \approx \mathbb{E}_{p(\theta|X)} [f(\theta)], \text{ with uncertainty proportional to } \sqrt{\text{Var}_{p(\theta|X)} (f(\theta))}.$$

**Practically:** We compute Markov chain Monte Carlo (MCMC) draws from  $p(\theta|X)$ .

# Economist 2016 Election Model [Gelman and Heidemanns, 2020]



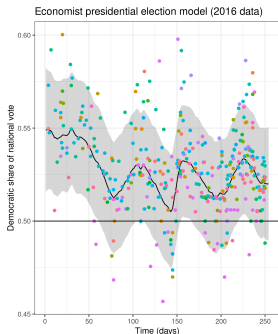
A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \dots, x_N =$  Polling data ( $N = 361$ ).
- $\theta =$  Lots of random effects (day, pollster, etc.)
- $f(\theta) =$  Democratic % of vote on election day

We want to know  $\mathbb{E}_{p(\theta|X)}[f(\theta)]$ .

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior  $p(\theta|X)$ .

# Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \dots, x_N =$  Polling data ( $N = 361$ ).
- $\theta =$  Lots of random effects (day, pollster, etc.)
- $f(\theta) =$  Democratic % of vote on election day

We want to know  $\mathbb{E}_{p(\theta|X)} [f(\theta)]$ .

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior  $p(\theta|X)$ .

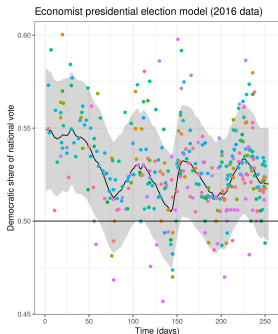
The people who responded to the polls were randomly selected.

If we had selected a different random sample, how much would our estimate have changed?

How can we estimate  $\text{Var}_{X \sim \mathbb{F}} \left( \mathbb{E}_{p(\theta|X)} [f(\theta)] \right)$ ?



# Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

- $X = x_1, \dots, x_N =$  Polling data ( $N = 361$ ).
- $\theta =$  Lots of random effects (day, pollster, etc.)
- $f(\theta) =$  Democratic % of vote on election day

We want to know  $\mathbb{E}_{p(\theta|X)} [f(\theta)]$ .

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior  $p(\theta|X)$ .

The people who responded to the polls were randomly selected.

If we had selected a different random sample, how much would our estimate have changed?

How can we estimate  $\text{Var}_{X \sim \mathbb{F}} \left( \mathbb{E}_{p(\theta|X)} [f(\theta)] \right)$ ?

- Except in special cases, we expect  $\text{Var}_{p(\theta|X)} (f(\theta)) \neq \text{Var}_{X \sim \mathbb{F}} \left( \mathbb{E}_{p(\theta|X)} [f(\theta)] \right)$
- We are interested re-sampling for *this* election, not a hypothetical future election

**Idea:** Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

**Problem:** Each MCMC run takes about 10 hours (Stan, six cores).

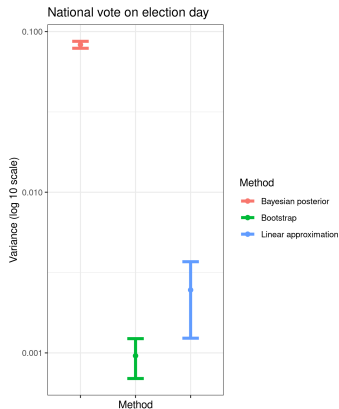
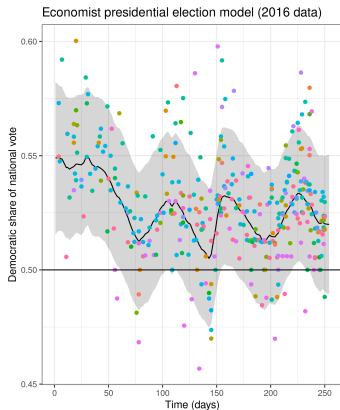
**Proposal:** Use full-data posterior draws to form a linear approximation to *data reweightings*.

# Results

**Idea:** Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

**Problem:** Each MCMC run takes about 10 hours (Stan, six cores).

**Proposal:** Use full-data posterior draws to form a linear approximation to *data reweightings*.

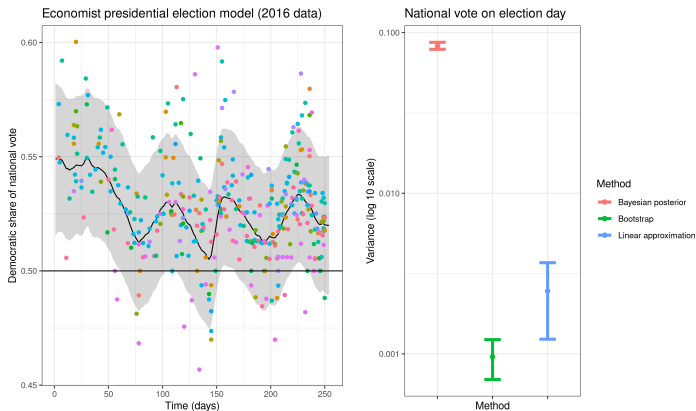


# Results

**Idea:** Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

**Problem:** Each MCMC run takes about 10 hours (Stan, six cores).

**Proposal:** Use full-data posterior draws to form a linear approximation to *data reweightings*.



Compute time for 100 bootstraps: 51 days

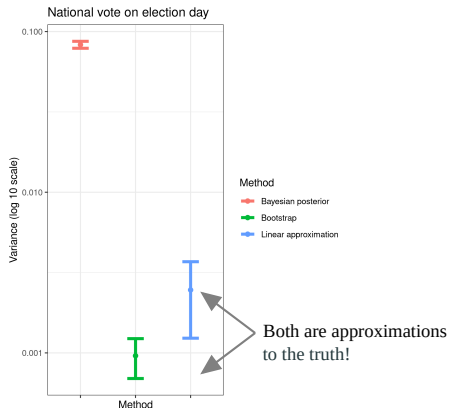
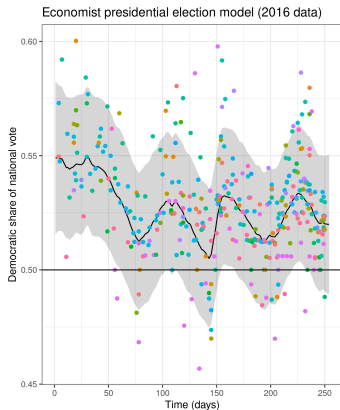
Compute time for the linear approximation: Seconds  
(But note the approximation has some error)

# Results

**Idea:** Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

**Problem:** Each MCMC run takes about 10 hours (Stan, six cores).

**Proposal:** Use full-data posterior draws to form a linear approximation to *data reweightings*.



Compute time for 100 bootstraps: 51 days

Compute time for the linear approximation: Seconds  
(But note the approximation has some error)

## Data re-weighting.

Augment the problem with *data weights*  $w_1, \dots, w_N$ .

## Data re-weighting.

Augment the problem with *data weights*  $w_1, \dots, w_N$ .

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta) \quad \Rightarrow \text{We write } \mathbb{E}_{p(\theta|X, w)} [f(\theta)].$$

Original weights:



## Data re-weighting.

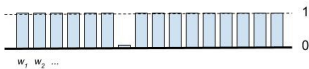
Augment the problem with *data weights*  $w_1, \dots, w_N$ .

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta) \quad \Rightarrow \text{We write } \mathbb{E}_{p(\theta|X, w)} [f(\theta)].$$

Original weights:



Leave-one-out weights:





# Data re-weighting.

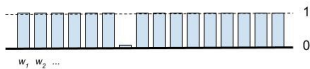
Augment the problem with *data weights*  $w_1, \dots, w_N$ .

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta) \quad \Rightarrow \text{We write } \mathbb{E}_{p(\theta|X, w)} [f(\theta)].$$

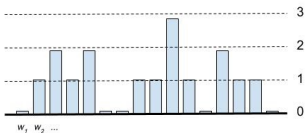
Original weights:



Leave-one-out weights:



Bootstrap weights:

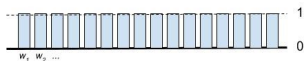


# Data re-weighting.

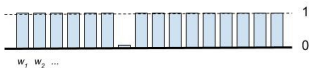
Augment the problem with *data weights*  $w_1, \dots, w_N$ .

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta) \quad \Rightarrow \text{We write } \mathbb{E}_{p(\theta|X, w)} [f(\theta)].$$

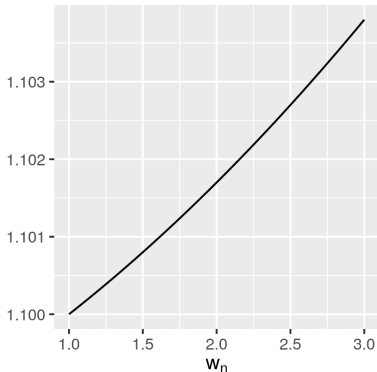
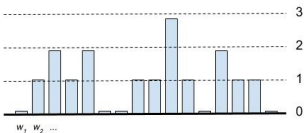
Original weights:



Leave-one-out weights:



Bootstrap weights:

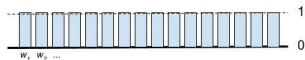


# Data re-weighting.

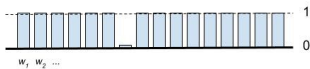
Augment the problem with *data weights*  $w_1, \dots, w_N$ .

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta) \quad \Rightarrow \text{We write } \mathbb{E}_{p(\theta|X, w)} [f(\theta)].$$

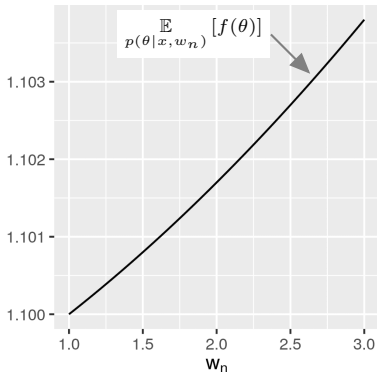
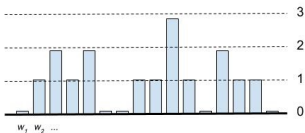
Original weights:



Leave-one-out weights:



Bootstrap weights:



# Data re-weighting.

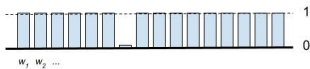
Augment the problem with *data weights*  $w_1, \dots, w_N$ .

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta) \quad \Rightarrow \text{We write } \mathbb{E}_{p(\theta|X, w)} [f(\theta)].$$

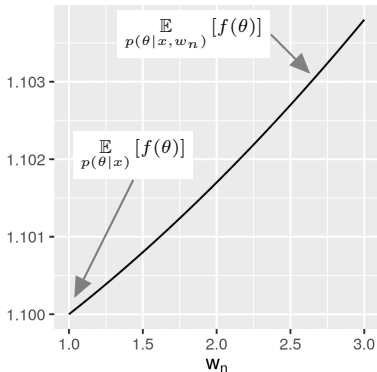
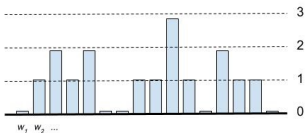
Original weights:



Leave-one-out weights:



Bootstrap weights:



# Data re-weighting.

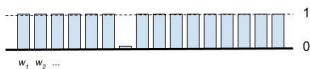
Augment the problem with *data weights*  $w_1, \dots, w_N$ .

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta) \quad \Rightarrow \text{We write } \mathbb{E}_{p(\theta|X, w)} [f(\theta)].$$

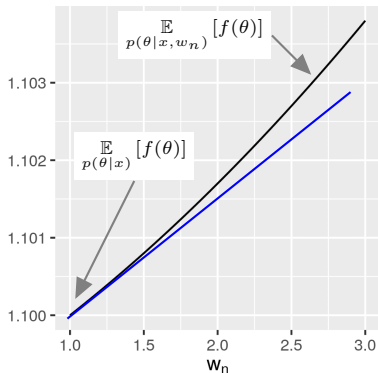
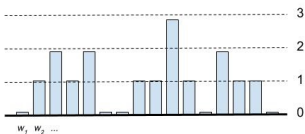
Original weights:



Leave-one-out weights:



Bootstrap weights:



# Data re-weighting.

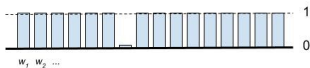
Augment the problem with *data weights*  $w_1, \dots, w_N$ .

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta) \quad \Rightarrow \text{We write } \mathbb{E}_{p(\theta|X, w)} [f(\theta)].$$

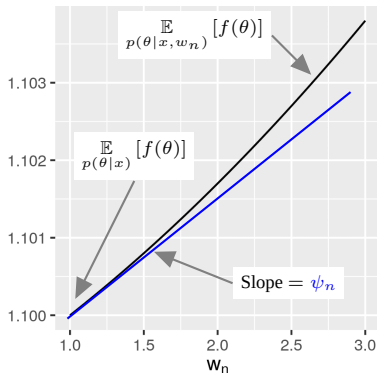
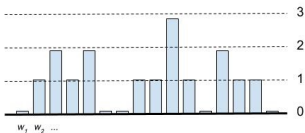
Original weights:



Leave-one-out weights:



Bootstrap weights:



# Data re-weighting.

Augment the problem with *data weights*  $w_1, \dots, w_N$ .

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta) \quad \Rightarrow \text{We write } \mathbb{E}_{p(\theta|X, w)} [f(\theta)].$$

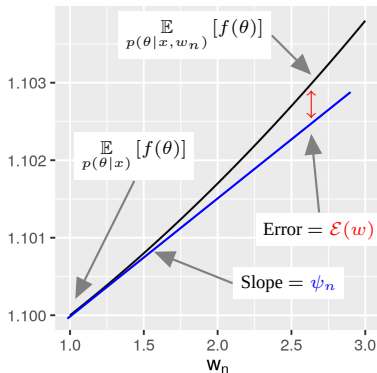
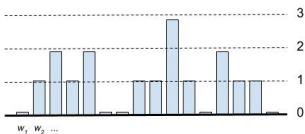
Original weights:



Leave-one-out weights:



Bootstrap weights:



# Data re-weighting.

Augment the problem with *data weights*  $w_1, \dots, w_N$ .

$$\ell_n(\theta) := \log p(x_n|\theta) \quad \log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta) \quad \Rightarrow \text{We write } \mathbb{E}_{p(\theta|X, w)} [f(\theta)].$$

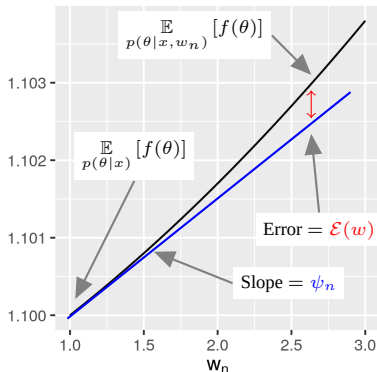
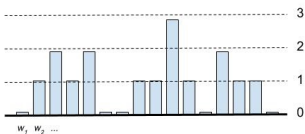
Original weights:



Leave-one-out weights:



Bootstrap weights:



The re-scaled slope  $N\psi_n$  is known as the “influence function” at data point  $x_n$ .

$$\mathbb{E}_{p(\theta|X, w)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] = \sum_{n=1}^N \psi_n (w_n - 1) + \mathcal{E}(w).$$



How can we use the approximation?

**Example: Approximate bootstrap.**

Draw bootstrap weights  $w \sim p(w) = \text{Multinomial}(N, N^{-1})$ .

This is equivalent to re-sampling data with replacement.

How can we use the approximation?

### Example: Approximate bootstrap.

Draw bootstrap weights  $w \sim p(w) = \text{Multinomial}(N, N^{-1})$ .

This is equivalent to re-sampling data with replacement.

$$\begin{aligned}\text{Bootstrap variance} &= \text{Var}_{p(w)} \left( \mathbb{E}_{p(\theta|x,w)} [f(\theta)] \right) \\ &\approx \text{Var}_{p(w)} \left( \mathbb{E}_{p(\theta|x)} [f(\theta)] + \psi_n(w_n - 1) \right) \quad (\text{assuming the error is small}) \\ &= \sum_{n=1}^N \left( \psi_n - \bar{\psi} \right)^2.\end{aligned}$$

The final line is also known as the “infinitesimal jackknife” variance approximation.

How can we use the approximation?

**Example: Approximate bootstrap.**

$$\text{Bootstrap variance} \approx \sum_{n=1}^N \left( \psi_n - \bar{\psi} \right)^2 .$$

How can we use the approximation?

**Example: Approximate bootstrap.**

$$\text{Bootstrap variance} \approx \sum_{n=1}^N \left( \psi_n - \bar{\psi} \right)^2.$$

**Example: Cross validation.** Let  $w_{(-n)}$  leave out point  $n$ , and loss  $f(\theta) = -\ell(x_n|\theta)$ .

$$\text{LOO CV loss at point } n = \mathbb{E}_{p(\theta|x, w_{(-n)})} [-\ell(x_n|\theta)] \approx \mathbb{E}_{p(\theta|x)} [-\ell(x_n|\theta)] - \psi_n$$

How can we use the approximation?

**Example: Approximate bootstrap.**

$$\text{Bootstrap variance} \approx \sum_{n=1}^N \left( \psi_n - \bar{\psi} \right)^2.$$

**Example: Cross validation.** Let  $w_{(-n)}$  leave out point  $n$ , and loss  $f(\theta) = -\ell(x_n|\theta)$ .

$$\text{LOO CV loss at point } n = \mathbb{E}_{p(\theta|x, w_{(-n)})} [-\ell(x_n|\theta)] \approx \mathbb{E}_{p(\theta|x)} [-\ell(x_n|\theta)] - \psi_n$$

**Example: Influential subsets: Approximate maximum influence perturbation (AMIP).**

Let  $W_{(-K)}$  denote weights leaving out  $K$  points.

$$\max_{w \in W_{(-K)}} \left( \mathbb{E}_{p(\theta|x, w)} [f(\theta)] - \mathbb{E}_{p(\theta|x)} [f(\theta)] \right) \approx - \sum_{n=1}^K \psi_{(n)}.$$

How can we use the approximation?

**Example: Approximate bootstrap.**

$$\text{Bootstrap variance} \approx \sum_{n=1}^N \left( \psi_n - \bar{\psi} \right)^2.$$

**Example: Cross validation.** Let  $w_{(-n)}$  leave out point  $n$ , and loss  $f(\theta) = -\ell(x_n|\theta)$ .

$$\text{LOO CV loss at point } n = \mathbb{E}_{p(\theta|x, w_{(-n)})} [-\ell(x_n|\theta)] \approx \mathbb{E}_{p(\theta|x)} [-\ell(x_n|\theta)] - \psi_n$$

**Example: Influential subsets: Approximate maximum influence perturbation (AMIP).**

Let  $W_{(-K)}$  denote weights leaving out  $K$  points.

$$\max_{w \in W_{(-K)}} \left( \mathbb{E}_{p(\theta|x, w)} [f(\theta)] - \mathbb{E}_{p(\theta|x)} [f(\theta)] \right) \approx - \sum_{n=1}^K \psi_n.$$

How to compute the slopes  $\psi_n$ ?

How large is the error  $\mathcal{E}(w)$ ?

## Expressions for the slope and error

How to compute the slopes  $\psi_n$ ? How can we analyze the error  $\mathcal{E}(w)$ ?

$$\mathbb{E}_{p(\theta|X,w)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] = \sum_{n=1}^N \psi_n (w_n - 1) + \mathcal{E}(w).$$

## Expressions for the slope and error

How to compute the slopes  $\psi_n$ ? How can we analyze the error  $\mathcal{E}(w)$ ?

$$\mathbb{E}_{p(\theta|X,w)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] = \sum_{n=1}^N \psi_n (w_n - 1) + \mathcal{E}(w).$$

By dominated convergence,  $\psi_n = \underbrace{\text{Cov}_{p(\theta|X)}(f(\theta), \ell_n(\theta))}_{\text{Estimatable with MCMC!}}.$



# Expressions for the slope and error

How to compute the slopes  $\psi_n$ ? How can we analyze the error  $\mathcal{E}(w)$ ?

$$\mathbb{E}_{p(\theta|X,w)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] = \sum_{n=1}^N \psi_n (w_n - 1) + \mathcal{E}(w).$$

By dominated convergence,  $\psi_n = \underbrace{\text{Cov}_{p(\theta|X)}(f(\theta), \ell_n(\theta))}_{\text{Estimatable with MCMC!}}$ .

Furthermore, by the mean value theorem, for some  $\tilde{w}$ ,

$$\mathcal{E}(w) = \frac{1}{2} \sum_{n=1}^N \sum_{n'=1}^N \mathcal{E}_{nn'}(w) (w_n - 1)(w_{n'} - 1) \quad \text{where}$$
$$\mathcal{E}_{nn'}(w) := \underbrace{\mathbb{E}_{p(\theta|X,\tilde{w})} [\bar{f}(\theta) \bar{\ell}_n(\theta) \bar{\ell}_{n'}(\theta)]}_{\substack{\text{Cannot compute directly!} \\ \text{(we don't know the intermediate value theorem's } \tilde{w} \text{).} \\ \text{But we can analyze it.}}}$$

Here, an overbar denotes “posterior–mean zero.” For example,  $\bar{f}(\theta) := f(\theta) - \mathbb{E}_{p(\theta|X)} [f(\theta)]$ .

How good is the linear approximation (IJ covariance) as an approximation of the limiting variance of  $\sqrt{N} \mathbb{E}_{p(\theta|X)} [f(\theta)]$ ?

# Theoretical results

How good is the linear approximation (IJ covariance) as an approximation of the limiting variance of  $\sqrt{N} \mathbb{E}_{p(\theta|X)} [f(\theta)]$ ?

**Theorem 3 of Giordano and Broderick [2024] (paraphrase):**

If the parameter dimension is fixed, and Bernstein–von Mises (BVM) theorem–like conditions hold, then the IJ covariance is consistent, because  $\sqrt{N}\mathcal{E}(w) \xrightarrow[N \rightarrow \infty]{prob} 0$ .

# Theoretical results

How good is the linear approximation (IJ covariance) as an approximation of the limiting variance of  $\sqrt{N} \mathbb{E}_{p(\theta|X)} [f(\theta)]$ ?

**Theorem 3 of Giordano and Broderick [2024] (paraphrase):**

If the parameter dimension is fixed, and Bernstein–von Mises (BVM) theorem–like conditions hold, then the IJ covariance is consistent, because  $\sqrt{N}\mathcal{E}(w) \xrightarrow[N \rightarrow \infty]{prob} 0$ .

Problem: we're doing MCMC because BVM does not hold.

What if  $f(\theta)$  concentrates marginally, but some components don't concentrate?

# Theoretical results

How good is the linear approximation (IJ covariance) as an approximation of the limiting variance of  $\sqrt{N} \mathbb{E}_{p(\theta|X)} [f(\theta)]$ ?

## **Theorem 3 of Giordano and Broderick [2024] (paraphrase):**

If the parameter dimension is fixed, and Bernstein–von Mises (BVM) theorem–like conditions hold, then the IJ covariance is consistent, because  $\sqrt{N}\mathcal{E}(w) \xrightarrow[N \rightarrow \infty]{prob} 0$ .

Problem: we’re doing MCMC because BVM does not hold.

What if  $f(\theta)$  concentrates marginally, but some components don’t concentrate?

## **Theorem 4 of Giordano and Broderick [2024] (paraphrase & conjecture):**

In a flexible class of high–dimensional exponential family models,

**even when  $p(f(\theta)|X)$  obeys a BVM marginally (!),**

- $\sqrt{N}\mathcal{E}(w)$  does not converge to zero (so the IJ covariance is inconsistent), but...
- $\sqrt{N}\mathcal{E}(w) = \tilde{O}_p(1)$ , and proportional to the nuisance parameters’ posterior covariance

# Theoretical results

How good is the linear approximation (IJ covariance) as an approximation of the limiting variance of  $\sqrt{N} \mathbb{E}_{p(\theta|X)} [f(\theta)]$ ?

## **Theorem 3 of Giordano and Broderick [2024] (paraphrase):**

If the parameter dimension is fixed, and Bernstein–von Mises (BVM) theorem–like conditions hold, then the IJ covariance is consistent, because  $\sqrt{N}\mathcal{E}(w) \xrightarrow[N \rightarrow \infty]{prob} 0$ .

Problem: we’re doing MCMC because BVM does not hold.

What if  $f(\theta)$  concentrates marginally, but some components don’t concentrate?

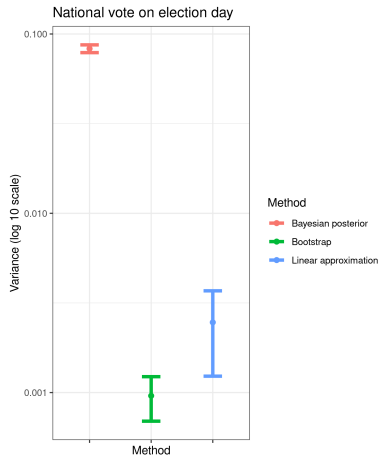
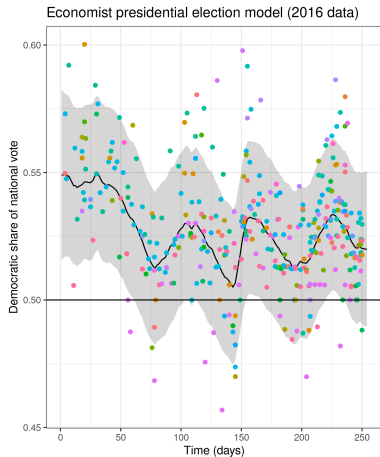
## **Theorem 4 of Giordano and Broderick [2024] (paraphrase & conjecture):**

In a flexible class of high–dimensional exponential family models, even when  $p(f(\theta)|X)$  obeys a BVM marginally (!),

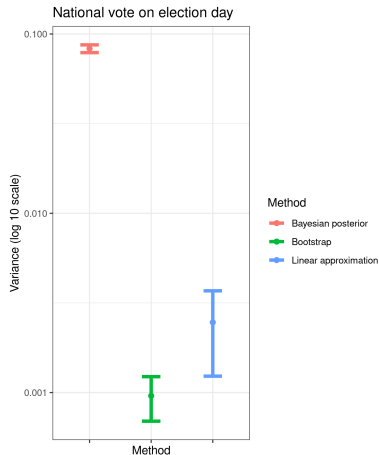
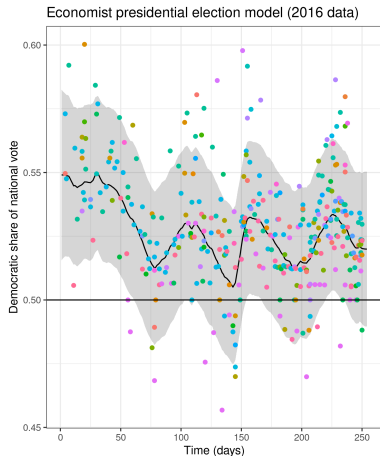
- $\sqrt{N}\mathcal{E}(w)$  does not converge to zero (so the IJ covariance is inconsistent), but...
- $\sqrt{N}\mathcal{E}(w) = \tilde{O}_p(1)$ , and proportional to the nuisance parameters’ posterior covariance

**$\Rightarrow$  High–dimensional Bayesian models are an extremely common class of problems for which the influence function may not provide a good approximation.**

# Observations and consequences



# Observations and consequences



**Preprint:** Giordano and Broderick [2024] (arXiv:2305.06466)

- Detailed proofs
- Simple analytical examples
- Simulated and real-world experiments



- A. Gelman and M. Heidemanns. The Economist: Forecasting the US elections., 2020. URL <https://projects.economist.com/us-2020-forecast/president>. Data and model accessed Oct., 2020.
- R. Giordano and T. Broderick. The Bayesian infinitesimal jackknife for variance. *arXiv preprint arXiv:2305.06466*, 2024.
- J. Huggins and J. Miller. Reproducible model selection using bagged posteriors. *Bayesian Analysis*, 18(1):79–104, 2023.

## Example: A negative binomial model

Consider  $p(X|\gamma) = \prod_{n=1}^N \text{NegativeBinomial}(x_n|\gamma)$ . Here,  $\theta = \gamma$  is a scalar.

## Example: A negative binomial model

Consider  $p(X|\gamma) = \prod_{n=1}^N \text{NegativeBinomial}(x_n|\gamma)$ . Here,  $\theta = \gamma$  is a scalar.

As  $N \rightarrow \infty$ ,  $p(\gamma|X)$  concentrates at rate  $1/\sqrt{N}$  (Bernstein–von Mises).

$$\Rightarrow N \left( \mathbb{E}_{p(\gamma|X, w_n)} [\gamma] - \mathbb{E}_{p(\gamma|X)} [\gamma] \right) = \psi_n(w_n - 1) + O_p(N^{-1}).$$

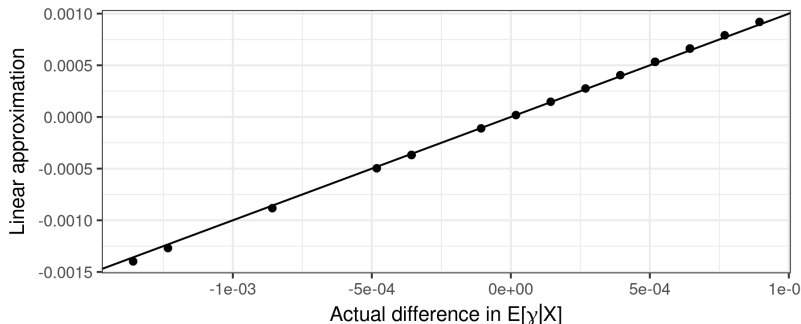
## Example: A negative binomial model

Consider  $p(X|\gamma) = \prod_{n=1}^N \text{NegativeBinomial}(x_n|\gamma)$ . Here,  $\theta = \gamma$  is a scalar.

As  $N \rightarrow \infty$ ,  $p(\gamma|X)$  concentrates at rate  $1/\sqrt{N}$  (Bernstein–von Mises).

$$\Rightarrow N \left( \mathbb{E}_{p(\gamma|X, w_n)}[\gamma] - \mathbb{E}_{p(\gamma|X)}[\gamma] \right) = \psi_n(w_n - 1) + O_p(N^{-1}).$$

Negative Binomial model  
leaving out single datapoints with  $N = 800$



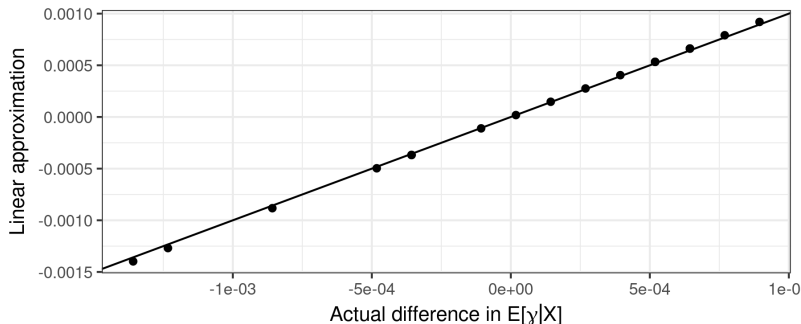
## Example: A negative binomial model

Consider  $p(X|\gamma) = \prod_{n=1}^N \text{NegativeBinomial}(x_n|\gamma)$ . Here,  $\theta = \gamma$  is a scalar.

As  $N \rightarrow \infty$ ,  $p(\gamma|X)$  concentrates at rate  $1/\sqrt{N}$  (Bernstein–von Mises).

$$\Rightarrow N \left( \mathbb{E}_{p(\gamma|X, w_n)}[\gamma] - \mathbb{E}_{p(\gamma|X)}[\gamma] \right) = \psi_n(w_n - 1) + O_p(N^{-1}).$$

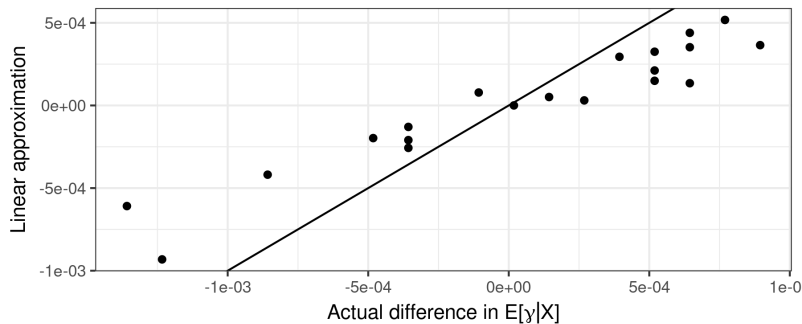
Negative Binomial model  
leaving out single datapoints with  $N = 800$



**Problem:** Most computationally hard Bayesian problems don't concentrate.

Example: **Poisson model with random effects (REs)  $\lambda$  and fixed effect  $\gamma$ .**

Poisson random effect model  
leaving out single datapoints with  $N = 800$



## A contradiction?

**Negative binomial observations.**

**Asymptotically linear in  $w$ .**

**Poisson observations with random effects.**

**Asymptotically non-linear in  $w$ .**

## A contradiction?

**Negative binomial observations.**

**Asymptotically linear in  $w$ .**

**Poisson observations with random effects.**

**Asymptotically non-linear in  $w$ .**

With a constant regressor, Gamma REs, and one RE per observation,  
these are the same model, with the same  $p(\gamma|X)$ .

**Is  $\mathbb{E}_{p(\gamma|X,w)} [\gamma]$  linear in the data weights or not?**



## A contradiction?

**Negative binomial observations.**

**Asymptotically linear in  $w$ .**

$$\log p(X|\gamma, w^m) = \sum_{n=1}^N w_n^m \log p(x_n|\gamma)$$

**Poisson observations with random effects.**

**Asymptotically non-linear in  $w$ .**

$$\log p(X|\gamma, \lambda, w^c) = \sum_{n=1}^N w_n^c \log p(x_n|\lambda, \gamma)$$

With a constant regressor, Gamma REs, and one RE per observation,  
these are the same model, with the same  $p(\gamma|X)$ .

Is  $\mathbb{E}_{p(\gamma|X, w)} [\gamma]$  **linear in the data weights** or not?

**Trick question!** We weight a log likelihood contribution, not a datapoint.

**The two weightings are not equivalent in general.**

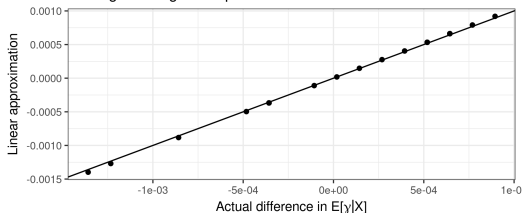
# Experimental results

Our results were actually computed on **identical datasets** with  $G = N$  and  $g_n = n$ .

Uses  $\log p(x_n | \gamma)$ :

$$\psi_n = \mathbb{E}_{p(\gamma|X)} [\bar{\gamma} \bar{\ell}_n(\gamma)]$$

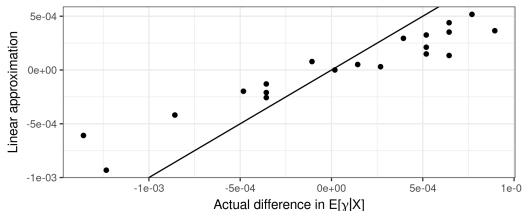
Negative Binomial model  
leaving out single datapoints with  $N = 800$



Uses  $\log p(x_n | \gamma, \lambda)$ :

$$\psi_n = \mathbb{E}_{p(\gamma, \lambda|X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)]$$

Poisson random effect model  
leaving out single datapoints with  $N = 800$



# Experimental results

Our results were actually computed on **identical datasets** with  $G = N$  and  $g_n = n$ .

Uses  $\log p(x_n | \gamma)$ :

$$\psi_n = \mathbb{E}_{p(\gamma|X)} [\bar{\gamma} \bar{\ell}_n(\gamma)]$$

Not computable from

$$\gamma, \lambda \sim p(\gamma, \lambda | X)$$

in general.

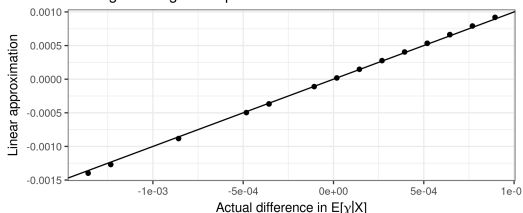
Uses  $\log p(x_n | \gamma, \lambda)$ :

$$\psi_n = \mathbb{E}_{p(\gamma, \lambda | X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)]$$

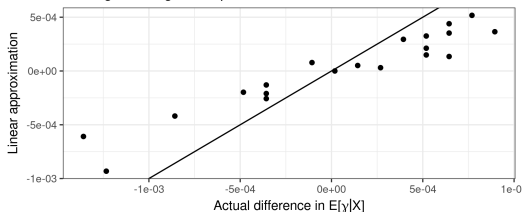
Computable from

$$\gamma, \lambda \sim p(\gamma, \lambda | X).$$

Negative Binomial model  
leaving out single datapoints with  $N = 800$



Poisson random effect model  
leaving out single datapoints with  $N = 800$



# Experimental results

Our results were actually computed on **identical datasets** with  $G = N$  and  $g_n = n$ .

Uses  $\log p(x_n | \gamma)$ :

$$\psi_n = \mathbb{E}_{p(\gamma|X)} [\bar{\gamma} \bar{\ell}_n(\gamma)]$$

Not computable from

$$\gamma, \lambda \sim p(\gamma, \lambda | X)$$

in general.

Uses  $\log p(x_n | \gamma, \lambda)$ :

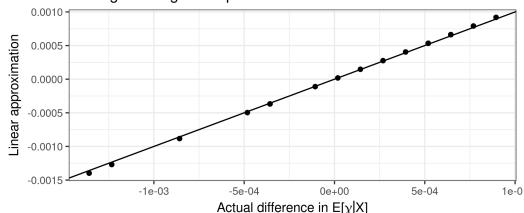
$$\psi_n = \mathbb{E}_{p(\gamma, \lambda | X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)]$$

Computable from

$$\gamma, \lambda \sim p(\gamma, \lambda | X).$$

May still be useful when  $p(\lambda | X)$  is *somewhat* concentrated.

Negative Binomial model  
leaving out single datapoints with  $N = 800$



Poisson random effect model  
leaving out single datapoints with  $N = 800$

